June the 6th, 2017

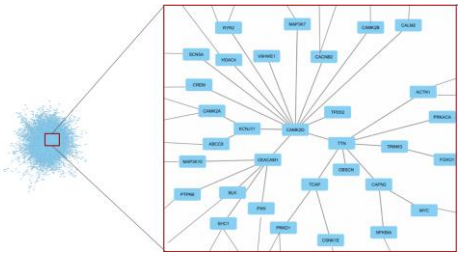
The facebook of genes

Community identification in biological networks

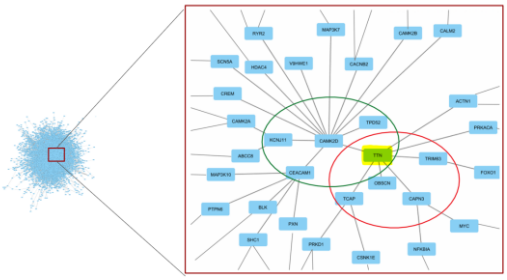
Marianne Bachmann, Philippe Fuchs and Thiery Masserey
With the supervision of Daniel Marbach

Introduction

In biology, many processes involve a group of genes that are expressed at the same time or a group of proteins that interact with each other. These interactions can be represented in a network where the nodes are the genes or proteins, and the links between the two nodes represent their relation.

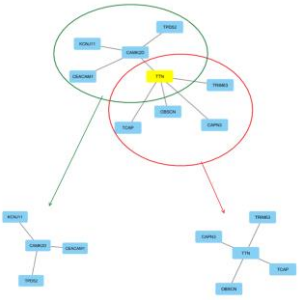


We had the goal to identify in this big networks some modules of genes, which are groups of densely interconnected nodes. Nodes which are part of the same module should have a related function or work in the same pathway. Thus, identifying modules allows a better understanding of biological processes, pathways and disease causes.



For our project, we wanted to first identify overlapping communities using a method known as *linkcomm*. An overlapping community is a community where nodes can be part of more than one community. It's particularly interesting in a biological context because one protein can be involved in different processes, as a node can be part of differents modules.

Then, we created an algorithm that allowed us to transform the overlapping communities into non-overlapping ones. In a non-overlapping community, one node can only be part of one community. Considering the network in terms of non-overlapping communities allows us to compare our results to the results of different methods.



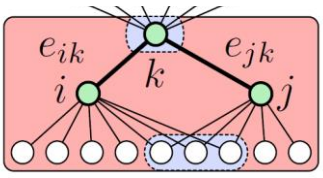
Methods

Linkcomm

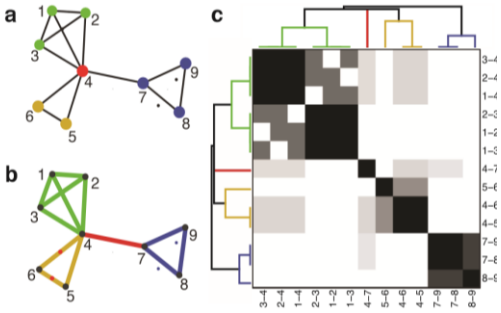
The linkcomm method allows to identify overlapping communities out of the network, which without any analysis almost resembles a hairball. To identify these communities, the algorithm uses a formula known as the Jacquard coefficient or index. It quantifies the similarity between the links: the stronger the similarity, the more likely they belong to the same community.

The Jacquard coefficient does not only consider the nodes connected by the links of interest, but also the neighbouring information. The more nodes are shared, the closer the links. For example, in the figure, edges e_{ik} and e_{jk} are related by four nodes (delimited by a violet area) out of a total of 12 nodes, resulting in a similarity of 0,3. The Jacquard index is comprised between 0 and 1.

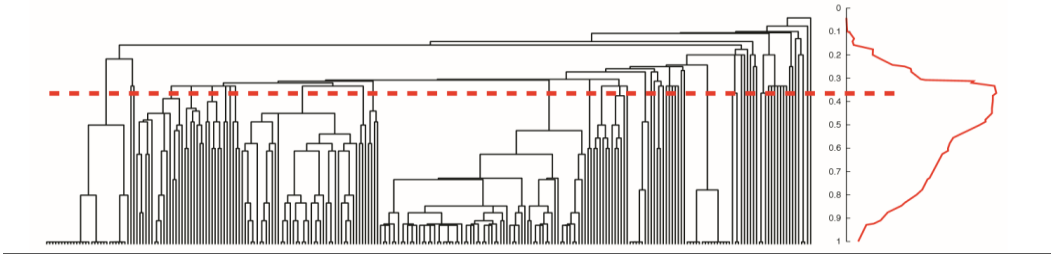
$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}$$



After analyzing the similarity between all the edges of the network, their Jacquard coefficients can be represented in a similarity matrix. In the figure on the right, the similarity is represented graphically: the darker the squares, the more similar are the nodes.



The following step is a single-linkage hierarchical clustering. At each iteration, the most similar edges are grouped together. The process is repeated until all edges are organized in a dendrogram, like the one shown the figure below.

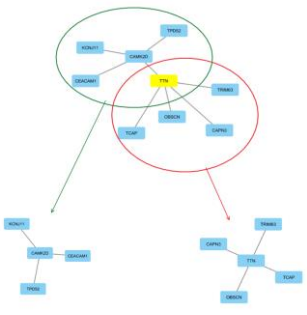


The dendrogram must be cut in order to identify the different communities. Cutting the tree is an important step: if cut too low, no relevant information can be extracted, since it would correspond to a list of unrelated links. When cut too high in the dendrogram, the edges would all belong to the same community. To know where to cut accurately the dendrogram, the linkcomm method uses the partition density, which measures the quality of the link partition. It considers the maximum and minimum edges possible in the community in relation to the actual number of edges present. This gives a partition density (represent as a red curve in the figure) and indicates where to cut the dendrogram (dotted red line), thus forming the communities.

Algorithm

After using the linkcomm method, we obtained overlapping communities, which corresponds to communities linked by nodes that are present in more than one community. However, to allow a comparison with other techniques of community identification, we needed to separate these communities and make them independent.

To perform this, we created an algorithm that relocates all the overlapping nodes into only one community. On the example (the figure on the right), the yellow overlapping node has been placed in the red community and deleted from the green one.



This replacement of the overlapping nodes is not randomly performed. The aim is to replace each overlapping node in the community in which it maximizes the most the link-density.

More precisely, the objective was to maximize the global link-density of all the communities. To perform this, the algorithm calculates for each overlapping node the difference of link-density of each community (in which the node is) with and without the node. Then, the node is replaced in the community in which this difference of link-density is maximized (i.e. deleted from the other communities). This methodology allows to consider the effect of keeping or losing a node on the link-density. At the end, we keep only the communities composed of at least 3 nodes.

Results

For the analysis, we used a signaling network conformed by 5254 nodes (or genes) and 21'826 links between them.

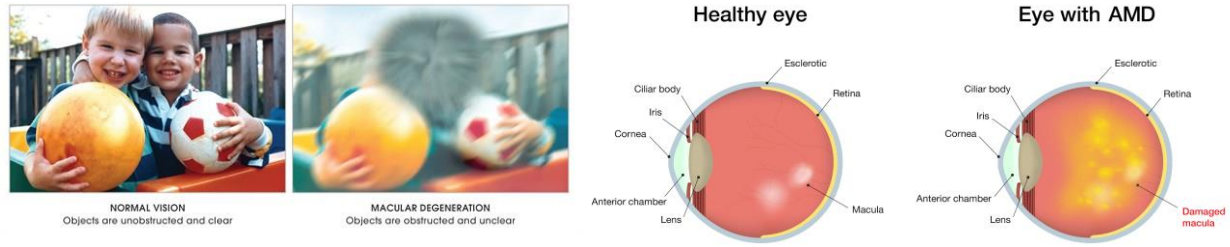
Linkcomm

Using the linkcomm method, we identified 880 communities in total. Out of the 5254 genes, 1190 weren't placed in a community, while 2'312 genes were placed in just one community and 1'752 belonged to more than one community, thus forming the overlaps.

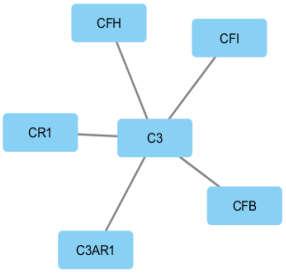
Algorithm

Using our algorithm, the 1752 overlapping nodes were each replaced in only one community. From the 880 communities extracted with the linkcomm method, only 416 communities remained. The 464 other communities, having less than 3 nodes, were deleted. Half of the remaining communities had only 3 or 4 nodes, while the biggest community was composed of 137 nodes.

Biological example: age-related macular degeneration



One of the communities we identified with the linkcomm method was composed of 6 genes that are all related to the age-related macular degeneration (AMD). It is a medical condition which results in blurred vision or no vision in the centre of the visual field. It typically occurs in older people, but genetic factors and smoking also play a role.



These 6 genes code for proteins that are all involved directly or indirectly to the complement system, which is part of the innate immune response.

This result allowed us to determine that the complement system plays a major role in the development of this disease.